

انتخاب ویژگی‌های بهینه به منظور کلاسه‌بندی داده‌های سرطانی مبتنی بر

الگوریتم بهینه‌سازی غیر جنسی

مرضیه محرمخانی^۱، علیرضا خان تیموری^۲، مجید مقدادی^۳

^۱ دانشجوی کارشناسی ارشد نرم افزار، گروه کامپیوتر، دانشگاه زنجان، زنجان
m.moharramkhani@znu.ac.ir

^۲ عضو هیئت علمی دانشکده مهندسی، گروه کامپیوتر، دانشگاه زنجان، زنجان
khanteymoori@znu.ac.ir

^۳ عضو هیئت علمی دانشکده مهندسی، گروه کامپیوتر، دانشگاه زنجان، زنجان
meghdadi@znu.ac.ir

چکیده

در این مقاله ایده‌ی جدیدی مبتنی بر الگوریتم بهینه‌سازی غیر جنسی برای جستجوی ویژگی‌های مناسب بر روی داده‌های سرطانی ارائه شده است. داده‌های سرطانی با ابعاد بالا، شامل تعداد زیادی ویژگی می‌باشند که اکثر این ویژگی‌ها غیر مرتبط با مسئله‌ی مورد نظر هستند. از طرف دیگر در داده‌های سرطانی، علاوه بر بالا بودن تعداد ویژگی‌ها، تعداد نمونه‌های موجود بسیار اندک است که این موضوع تاثیر نامناسبی بر روی کارایی الگوریتم‌های کلاسه‌بندی دارد. برای حل این مشکل، در این مقاله، الگوریتم جدیدی برای انتخاب ویژگی‌های بهینه مبتنی بر الگوریتم بهینه‌سازی غیر جنسی معرفی شده است.

الگوریتم بهینه‌سازی غیر جنسی یکی از الگوریتم‌های تکاملی الهام گرفته از زندگی موجودات تک جنسی است. این الگوریتم به دلیل عدم نیاز به تنظیم پارامتر و عدم نیاز به عملگر انتخاب در مقایسه با الگوریتم ژنتیک بسیار سریع‌تر عمل می‌کند و از افتادن در بهینه محلی اجتناب می‌کند. الگوریتم ارائه شده بر روی دو مجموعه داده‌ی Colon و Tumor_9 تست شده است. قبل از عمل جستجو، به منظور عدم انتخاب ویژگی‌های افزونه از گروه بندی ویژگی‌ها استفاده شده است. پیاده سازی روش ارائه شده نشان می‌دهد که الگوریتم ARO در مقایسه با الگوریتم ژنتیک دارای سرعت و دقت بالایی در پیدا کردن نتیجه مطلوب است.

کلمات کلیدی

انتخاب ویژگی، الگوریتم ARO، داده‌های میکروآرایه، کلاسه بندی داده‌ها، الگوریتم ژنتیک، داده کاوی

وجود نویز در داده‌های میکروآرایه‌ها غیر قابل اجتناب است. برای حل این مشکلات روش‌های انتخاب ویژگی و استخراج ویژگی استفاده می‌شود.

هدف از انتخاب ویژگی، انتخاب یک مجموعه کمینه از ویژگی‌ها است، به گونه‌ای که دقت الگوریتم کلاسه بندی افزایش یابد [5]. انتخاب ویژگی علاوه بر حوزه تشخیص پزشکی می‌تواند در سایر زمینه‌های داده کاوی مثل پردازش تصویر، کلاسه بندی متن و ... نیز مورد استفاده قرار گیرد.

تاکنون محققان الگوریتم‌های زیادی برای انتخاب ویژگی‌های مناسب در داده‌های بیان ژن ارائه داده‌اند. به طور کلی می‌توان الگوریتم‌های انتخاب ویژگی را به سه دسته الگوریتم‌های Filter، الگوریتم‌های Wrapper و الگوریتم‌های Embedded تقسیم نمود [6]. الگوریتم‌های Filter، معمولاً از یک معیار ارزیابی به منظور بدست آوردن ارتباط بین ویژگی‌ها با برچسب کلاس استفاده می‌کنند و برای ویژگی‌ها بر اساس میزان ارتباطشان امتیازی در نظر می‌گیرند. بعد از

۱- مقدمه

روش‌های یادگیری ماشین بر روی مجموعه ای از داده‌ها کار می‌کنند که اغلب دارای ویژگی‌های زیادی می‌باشند. برای مثالی می‌توان به مجموعه داده‌های میکروآرایه‌ها اشاره نمود. میکروآرایه تکنیکی است که امکان بررسی سطح بیان هزاران ژن را در آن واحد فراهم می‌کند [3]. با وجود این که کلاسه بندی داده‌ها در حوزه داده کاوی، موضوع جدیدی محسوب نمی‌شود ولی با ظهور مجموعه داده‌هایی با ابعاد بالا در این حوزه، مشکلات جدیدی برای الگوریتم‌های کلاسه‌بندی به وجود آمده است. در داده‌های بیان ژن که در قالب میکروآرایه‌ها قابل دسترسی است، تعداد ویژگی‌ها یا ژن‌ها بسیار بالا و در مقابل تعداد نمونه‌های موجود بسیار اندک است. این عامل باعث می‌شود که الگوریتم‌های کلاسه بندی در مرحله آموزش، با تعداد نمونه‌های کافی آموزش نبینند و در نتیجه در مرحله تشخیص، نتوانند نمونه‌های جدید را به خوبی از هم متمایز کنند [4]. از طرف دیگر،